Weighted random generation of context-free languages: Analysis of collisions in random urn occupancy models

Danièle Gardy, Yann Ponty

WEIGHTED RANDOM GENERATION OF CONTEXT-FREE LANGUAGES: ANALYSIS OF COLLISIONS IN RANDOM URN OCCUPANCY MODELS

DANIÈLE GARDY AND YANN PONTY

Abstract. The present work analyzes the redundancy of sets of combinatorial objects produced by a weighted random generation algorithm proposed by Denise *et al.* This scheme associates weights to the terminals symbols of a weighted context-free grammar, extends this weight definition multiplicatively on words, and draws words of length n with probability proportional their weight. We investigate the level of redundancy within a sample of k word, the proportion of the total probability covered by k words (coverage), the time (number of generations) of the first collision, and the time of the full collection. For these four questions, we use an analytic urn analogy to derive asymptotic estimates and/or polynomially computable exact forms. We illustrate these tools by an analysis of an RNA secondary structure statistical sampling algorithm introduced by Ding *et al.*

1. Introduction

The random generation of combinatorial objects is both motivated by the exploration of complex objects, the empirical assessment of statistical properties and by its applications to numerous fields (analysis of data structures and algorithms [1], software testing [6, 5], bioinformatics [9]...). Many approaches have been developed to address the uniform random generation of combinatorial objects of a given size. Historically, the recursive method, formalized by Wilf [24], starts by efficiently precomputing the numbers of objects accessible from local choices, and uses these numbers during the generation to perform an uniform random generation as an unbiased walk. This approach was later extended and made fully automatic by Flajolet *et al* [15] for all decomposable combinatorial classes, i. e. classes that are specified constructively within the symbolic framework as opposed to implicitly defined by a required property. Finally Duchon *et al* [11] recently relaxed this scheme through Boltzmann sampling.

Yet certain contexts require a non-uniform – yet controlled – distribution to be captured, giving rise to various approaches [4] for the non-uniform generation. Denise *et al* [7] introduced weighted context-free grammars where a weight function, defined on the terminals and extended multiplicatively on words, induces a Boltzmann distribution over each subset of words of a given length n. The resulting languages are then used as models for objects following non-uniform distributions, of which natural instances can be found in bioinformatics [21]. An adaptation of the recursive method was proposed [7] to draw words of a given size n with respect to a weighted distribution. Multidimensional Boltzmann versions of the weighted samplers were also proposed for weighted languages by Bodini *et al* [3].

However weighted distributions, by assigning probabilities to possible words that scale exponentially within a class of size, may induce a – possibly large – redundancy within sampled set of words. Since the probability of a word is exactly and efficiently computable such a redundancy is not informative and should be avoided. Furthermore, if a non-redundant sample of given cardinality k is expected, one may find situations where the complexity of generating k distinct words using a rejection approach becomes heavily dominated by the rejection step. Finally, the proportion of the

Date: December 6, 2010.

Key words and phrases. Random generation, occupancy analysis, birthday paradox, coupon collector, weighted combinatorial objects.

distribution contained within a sampled set may be affected, positively or negatively, by the adjunction of weights. One of the authors proposed a non-redundant version of the recursive method [20] to work around the first issue. However the question of the dependency between the weights and the level of redundancy was left open in a general setting.

The aim of the current work is to analyze the redundancy and coverage of a weighted sampled set of words. To tackle these questions, one can reformulate the repeated generation of words within a weighted language as a random allocation of balls into urns. Namely each word w in $\mathcal{L}_n$ the restriction of the language to words of length n will correspond to an urn having probability proportional to the weight of w. A list of questions naturally arise which can be rephrased into classic random allocations problems:

(1) How many words are required before some word is drawn twice? This is a weighted instance of the Birthday *paradox* (the first 2-birthday [13]).
(2) How many words must be sampled before each word in $\mathcal{L}_n$ is encountered at least once? One finds in the above formulation the Coupon collector problem.
(3) How many distinct words are there after sampling k words? This is equivalent to the expected number of urns having positive load after throwing k balls.
(4) What is the coverage, i.e. the cumulated weight/probability of a non-redundant sampled set after k generations? This last problem rephrases as the cumulated weight/probability of urns having positive load after throwing k balls.

In this paper, we address and provide closed formulae and/or asymptotic estimates for these four statistical quantities under natural conditions of non-degeneracy, and illustrate our results with an analysis of a statistical sampling algorithm used to predict the folding of RNA. After this short introduction we remind in Section 2 some basic notions related to context-free grammars, languages, algebraic functions and their weighted analogs. In Section 3, we state our main results on weighted context-free languages in the form of four theorems dedicated to the four questions above. General results on weighted urns models are established or recalled in Section 4, of which our theorems are direct corollaries. We apply in Section 5 our theorems to an analysis of a statistical sampling algorithm used to predict the functional folding of RNAs, using the fact that the three-dimensional structure of an RNA can be modeled by a secondary structure, i. e. a word of a Motzkin-like context-free language. We conclude with some possible extensions of the current work.

2. Definitions and notations

2.1. **Weighted context-free languages.** Throughout the rest of the document, n will stand for the length of generated words. For the sake of self-containment, let us start by recalling some definitions found in Denise *et al* [7].

A **weighted context-free grammar** $\mathcal{G}_\pi$ is a 5-tuple $(\pi, \Sigma, \mathcal{N}, \mathcal{P}, \mathcal{S})$ such that

- Σ is the alphabet, i.e. a finite set of terminal symbols.
- $\mathcal{N}$ is a finite set of non-terminal symbols.
- $\mathcal{P}$ is the finite set of production rules, each of the form $N \to X$, for $N \in \mathcal{N}$ any non-terminal and $X \in \{\Sigma \cup \mathcal{N}\}^*$.
- $\mathcal{S}$ is the **axiom** of the grammar, i. e. the initial non-terminal.
- π is a positive **weight** vector $\pi = (\pi_t)_{t\in\Sigma}$, assigning positive weights to each letter $t_i \in \Sigma$.

Let us further assume that the input grammar is unambiguous. This is a real limitation, however a similar analysis for intrinsically ambiguous languages is rather challenging since the associated generating functions are not necessarily algebraic but possibly transcendental [12].

Let us denote by $\mathcal{L}$ be the language generated from the axiom of $\mathcal{G}_\pi$, and by $\mathcal{L}_n$ its restriction to words of size n. One can extend the weight multiplicatively on any word $w \in \mathcal{L}$ such that

$$\pi(w) = \prod_{t\in w} \pi_t.$$

This gives rise to the notion of **weighted generating function** $L_\pi(z)$ for a context-free language $\mathcal{L}$, a natural generalization of the ordinary generating function where each word is counted with multiplicity equal to its weight

$$L_\pi(z) = \sum_{w\in\mathcal{L}} \pi(w) z^{|w|} = \sum_{n\geq 0} \mu_{\pi,n} z^n$$

where $\mu_{n,\pi} = \sum_{w\in\mathcal{L}_n} \pi(w)$ is the total weight of words of size n. In particular, the number m_n of words of size n can be also expressed as $m_n = |\mathcal{L}_n| = \mu_{n,\mathbf{1}}$.

The weighting scheme then defines a **weighted distribution** on $\mathcal{L}_n$ through

$$\mathbb{P}(w \mid n, \pi) = \frac{\pi(w)}{\sum_{w'\in\mathcal{L}_n} \pi(w')} = \frac{\pi(w)}{\mu_{n,\pi}}.$$

Finally let us define the k**-th moment** of a π-weighted distribution as

$$\alpha_{k,n} = \sum_{i=1}^{m_n} p_i^k = \frac{\sum_{w\in\mathcal{L}_n} \pi(w)^k}{{\mu_{\pi,n}}^k} = \frac{\mu_{\pi^k,n}}{{\mu_{\pi,n}}^k}. \tag{2.1}$$

2.2. **Asymptotics of coefficients.** The (weighted) generating function of an unambiguous context-free language is a positive solution of an algebraic system of equations, therefore its singularities are algebraic. Let us first assume that the dominant singularity ρ_π is unique.

Then, for any fixed π, the coefficients of $L_\pi(z)$ admit an asymptotic equivalent of the form

$$[z^n]\, L_\pi(z) = \mu_{\pi,n} \sim \kappa_\pi \cdot \rho_\pi^{-n} \cdot n^{-k_\pi} \left(1 + \mathcal{O}(n^{-k'_\pi})\right), \tag{2.2}$$

for $\rho_\pi \in (0, 1]$, κ_π some positive real value, and k_π, k'_π some positive rational numbers such that $k'_\pi > 0$. The asymptotic equivalent for the number of words $m_n = |\mathcal{L}_n| = [z^n]\, L(z)$ is obtained as a special case of the above, yielding

$$m_n = |\mathcal{L}_n| = [z^n]\, L(z) \sim \kappa \cdot \rho^{-n} \cdot n^{-k} \left(1 + \mathcal{O}(n^{-k'})\right) \tag{2.3}$$

with $\rho := \rho_{\mathbf{1}}$, $\kappa := \kappa_{\mathbf{1}}$, $k := k_{\mathbf{1}}$ and $k' := k'_{\mathbf{1}} > 0$ defined as above.

If the assumption on the unicity of the dominant singularity does not hold, then different singularities may be found on the circle of radius ρ_π. In this case the coefficients of the generating functions do not admit an universal expansion of the form described in Equation 2.2 since the contributions of various singularities may cancel out.

2.3. **Weight classes.** Let us denote by $\mathbf{W}_n$ the vector of all possible and distinct weights within $\mathcal{L}_n$ ordered increasingly ($W_{n,i} < W_{n,i+1}$). In particular, let $W^{\nabla}_{\pi,n} := W_{n,1}$ (resp. $W^{\triangle}_{\pi,n} := W_{n,|\mathbf{W}_n|}$) be the **minimal** (resp. **maximal**) **weight** of a word within $\mathcal{L}_n$. We denote by $\mathbf{m}_{n,i} \subset \mathcal{L}_n$ the class of words having weight $W_{n,i}$ and by $m_{n,i} = |\mathbf{m}_{n,i}|$ its cardinality.

3. Main results

Let $\mathcal{G}_\pi$ be a weighted context-free grammar generating a language $\mathcal{L}$, π its a weight vector and $n \in \mathbb{N}$ a length. Remind that $W^{\nabla}_{\pi,n}$ and $W^{\triangle}_{\pi,n}$ are the minimal and maximal weight of a word in $\mathcal{L}_n$ respectively. Let ρ_π be the dominant singularity of $L_\pi(z)$, and consider the following conditions:

- **C1** Diversity: Let $p^{\triangle}_{n,\pi} := W^{\triangle}_{\pi,n}/\mu_{\pi,n}$ be the probability of the most probable word within $\mathcal{L}_n$ with respect to a weight function π, then there exists $\beta > 1$ such that $p^{\triangle}_{\pi,n} \in \mathcal{O}(\beta^{-n})$.
- **C2** Log-positive weights: For each terminal symbol $t \in \Sigma$, $\pi_t > 1$.
- **C3** Bounded dependency: For any rational number $k > 1$ and any weight vector π such that Condition **C2** holds, ${\rho_\pi}^k < \rho_{\pi^k}$ holds.

Theorem 3.1 (First collision)**.** *Under conditions* **C1**, **C2** *and* **C3**, *the expected number of generations* $E[B_{n,\pi}]$ *before some word of* $\mathcal{L}_n$ *is drawn twice is such that*

$$E[B_{n,\pi}] \sim \frac{\sqrt{\pi}}{\sqrt{2\alpha_{2,n}}} = \frac{\mu_{\pi,n}\sqrt{\pi}}{\sqrt{2\mu_{\pi^2,n}}} \in \Omega\left(\gamma^n\right), \quad \gamma := \frac{\rho_\pi}{\sqrt{\rho_{\pi^2}}} > 1 \tag{3.1}$$

Theorem 3.2 (Full collection)**.** *The expected number of generations* $E[C_{n,\pi}]$ *before all the words in* $\mathcal{L}_n$ *are generated at least once is such that*

$$\frac{\mu_{\pi,n}}{W^{\nabla}_{\pi,n}} \leq E[C_{n,\pi}] \leq 2 \cdot \mathcal{H}_{m_n} \cdot \frac{\mu_{\pi,n}}{W^{\nabla}_{\pi,n}} \tag{3.2}$$

which, for large values of n, *adopts the equivalent*

$$\frac{\kappa_\pi \cdot \rho_\pi^{-n}}{W^{\nabla}_{\pi,n} \cdot n^{k_\pi}} \leq E[C_{n,\pi}] \leq \frac{2 \cdot \log(1/\rho) \cdot \kappa_\pi \cdot \rho_\pi^{-n}}{W^{\nabla}_{\pi,n} \cdot n^{k_\pi - 1}}. \tag{3.3}$$

Moreover in the uniform distribution ($\pi = \mathbf{1}$) *the above expression simplifies into*

$$E[C_{n,1}] = m_n \cdot \mathcal{H}_{m_n} \sim \frac{\kappa \cdot \log(1/\rho) \cdot \rho^{-n}}{n^{k-1}} \left(1 + \mathcal{O}\left(1/n^{k'}\right)\right). \tag{3.4}$$

Theorem 3.3 (Distinct samples)**.** *The expected number* $E[N_{n,\pi,k}]$ *of distinct words obtained after* k *generations is such that*

$$E[N_{n,\pi,k}] = \sum_{i=1}^{|\mathbf{W}|} m_{\pi,i} \cdot \left(1 - \left(1 - \frac{W_{n,i}}{\mu_{\pi,n}}\right)^k\right) = \sum_{i=1}^{m} m_{\pi,i} \cdot \left(1 - e^{-\frac{W_{n,i}}{\mu_{\pi,n}}k}\right) + \mathcal{O}(1). \tag{3.5}$$

Theorem 3.4 (Coverage)**.** *In a weighted distribution, the expected cumulated probability* $E[P_{n,\pi,k}] \in [0,1]$ *of the set of distinct words obtained after* k *generations is given by*

$$E[P_{n,\pi,k}] = \sum_{i=1}^{|\mathbf{W}|} m_{\pi,i} \cdot \frac{W_{n,i}}{\mu_{\pi,n}} \cdot \left(1 - \left(1 - \frac{W_{n,i}}{\mu_{\pi,n}}\right)^k\right). \tag{3.6}$$

Moreover if Condition **C1** *is satisfied, then there exists* $\beta > 1$ *such that, for any* $k \in o(\beta^n)$, *one has*

$$E[P_{n,\pi,k}] = k \cdot \alpha_{2,n} \left(1 + \mathcal{O}(\beta^{-n})\right). \tag{3.7}$$

Remark that there are at most $(n+1)^{|\Sigma|}$ different compositions/classes of weights, and therefore Theorems 3.3 and 3.4 immediately suggest polynomial time algorithms for computing the expected number of distinct words and coverage respectively.

3.1. **Discussing the loss of generality.** Let us discuss the loss of generality induced by the above conditions:

- Condition **C1** requires that no polynomial group of words contribute asymptotically to a significant part of the weighted distribution. This is the typical case in weighted languages, as the exponential growth of $\mu_{\pi,n}$ usually arises as a cooperation between the natural combinatorial explosion of the numbers of words and their individual weights. However this condition is restrictive, and discards languages of polynomial growth, or grammars where a (strongly connected) component of polynomial growth dominates asymptotically.
- Condition **C2** can be assumed without loss of generality since the weighted distribution is stable through the multiplication of all weights by a positive constant.
- Condition **C3**: Remember that Condition **C1** implies that there exist some constants $C > 0$ and $\beta > 1$ such that $\pi(w) \leq C \cdot \mu_{\pi,n}/\beta^n$ for all $w \in \mathcal{L}_n$. It follows that, for all $k > 1$,

$$\mu_{\pi^k,n} = \sum_{w \in \mathcal{L}_n} \pi(w)^k \leq \sum_{w \in \mathcal{L}_n} \pi(w) \cdot \left(\frac{C \cdot \mu_{\pi,n}}{\beta^n}\right)^{k-1} = {\mu_{\pi,n}}^k \cdot \frac{C^{k-1}}{\beta^{(k-1)\cdot n}}.$$

Consequently the exponential growth factor $\rho_{\pi^k}^{-1}$ of $\mu_{\pi^k,n}$ is such that

$$\rho_{\pi^k}^{-1} \leq \left(\beta^{(k-1)}\rho_\pi^k\right)^{-1} < {\rho_\pi^k}^{-1}$$

and Condition **C3** is a direct consequence of Condition **C1**.

4. General theorems

In the following section we establish general results on non-uniform urn models, which we apply to weighted distributions. Let $\mathbf{u}$ be a set of urns, $m = |\mathbf{u}|$ its cardinality and, for each $u_i \in \mathbf{u}$, let W_i be the weight of u_i, and p_i its probability. This defines a probability distribution $\mathbf{p} = (p_i)_{i=1}^m$ such that $\sum_{i=1}^m p_i = 1$ and for all $i \in [1, m-1]$, $p_i \leq p_{i+1}$.

4.1. Birthday paradox: First collision.

Theorem 4.1. *Assume there exists $\tau := \tau(\mathbf{p})$ such that*

(A) $p_m \cdot \tau < 1$;
(B) $\sqrt{\alpha_2} \cdot \tau \to +\infty$ when $m \to \infty$;
(C) $\sqrt[3]{\alpha_3} \cdot \tau \to 0$ when $m \to \infty$;

Then the waiting time $E(B)$ of the first birthday can be approximated by

$$E(B) = \sqrt{\frac{\pi}{2\alpha_2}}(1 + o(1)).$$

4.1.1. *Application to weighted distribution.*

Proposition 4.2. *Let $\mathcal{G}_\pi$ be a weighted context-free grammar and $\mathcal{L}$ be its associated language, satisfying Conditions* **C1**, **C2**, *and* **C3**. *Then the weighted distribution induced on $\mathcal{L}_n$ satisfies the conditions (A), (B) and (C) of Theorem 4.1 for any $\tau_n := \alpha_{k,n}$ such that $2 < k < 3$. Consequently the first collision is observed after $E[B \mid n] = \sqrt{\pi/2\alpha_{2,n}}(1 + o(1))$ generations.*

4.2. Coupon collector: Waiting for the full collection.

First let us remind that the uniform case is covered by the following *folklore* theorem [13].

Theorem 4.3. *In the uniform distribution, the waiting time $E[C_{\mathbf{1}}]$ is given by*

$$E[C_{\mathbf{1}}] = m \cdot \mathcal{H}_m \in \Theta(m \cdot \log(m)). \tag{4.1}$$

Theorem 4.4. *In a non-uniform distribution and for large values of n, the waiting time $E[C_\pi]$ of the full collection obeys*

$$\frac{1}{p_1} \leq E[C_\pi] \leq 2 \cdot \mathcal{H}_m \cdot \frac{1}{p_1} \tag{4.2}$$

where p_1 is the smallest probability of an urn.

Proof. First let us point out that, for any urn u, the waiting time of the full collection is greater than the expected time when a first ball reaches u. Since the least probable urn has probability p_1, then the lower bound on $E[C_\pi]$ immediately follows.

From a recent contribution by Berenbrink and Sauerwald [2], we know that the waiting time $E[C_\pi]$ for the full collection of m items drawn with respective probabilities $p_1 \leq p_2 \leq \ldots \leq p_m$ can be approximated within a $\mathcal{O}(\log\log m)$ factor by an estimate

$$\mathcal{U}_m = \sum_{i=1}^{m} \frac{1}{ip_i}. \tag{4.3}$$

More precisely it is shown in [2] that

$$\frac{\mathcal{U}_m}{3e \cdot \log\log m} \leq E[C_\pi] \leq 2\mathcal{U}_m. \tag{4.4}$$

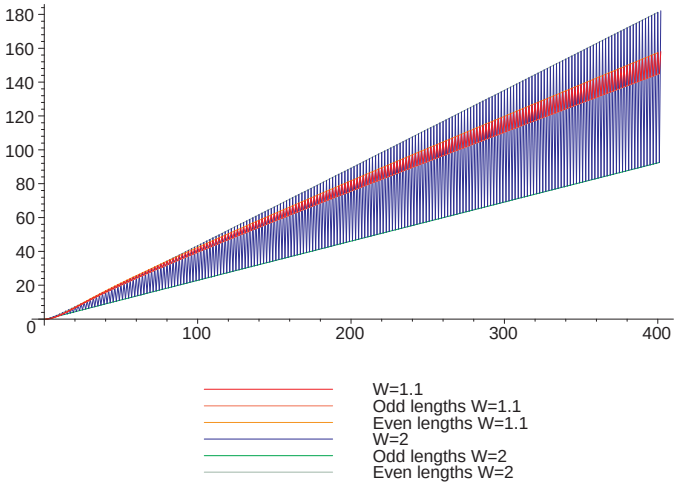

FIGURE 1. Plots of $p_1 \cdot \mathcal{U}_m$ for weighted Motzkin words exhibit a linear growth on n, suggesting that the upper bound is reached.

In our urn model, equation 4.3 specializes into

$$\mathcal{U}_m = \sum_{i=1}^{m} \frac{1}{ip_i} = \frac{1}{p_1}\left(\sum_{i=1}^{m} \frac{1}{i\Delta_i}\right)$$

where $\Delta_i := p_i/p_1$. Since p_1 is the weight of the least probable urn, then one has $\Delta_i \geq 1,\ \forall i \in [1, m]$, and therefore the following upper bound holds

$$\mathcal{U}_m \leq \frac{1}{p_1}\left(\sum_{i=1}^{m} \frac{1}{i}\right) = \frac{1}{p_1} \cdot \mathcal{H}_m$$

in which one recognizes the upper bound of Equation 4.2. □

Experiments suggest that the upper bound is in fact reached. For instance, Figure 1 shows the value $p_1 \cdot \mathcal{U}_m$ for weighted Motzkin paths, where a weight $W > 1$ is associated to horizontal steps, while up and down steps remain unweighted. In such a case the growth of $p_1 \cdot \mathcal{U}_m$ appears to be linear with different slopes depending on the parity of n. This phenomenon is due to the fact that the minimal number of horizontal steps in a Motzkin word of length n is 0 (resp. 1) for even (resp. odd) lengths, leading to minimal weights of 1 for even lengths and π to odd ones.

4.3. **Occupancy analysis.** Figuring out the average state after k generations turns out to be easier that the inverse problem – finding expected number k of generations before a given state is observed. We refer to a survey [16] by one of the authors for examples of urns model in the context of the analysis of algorithms. Here we establish a general formula for the cumulated weight in a weighted urn model through a generating function analysis.

Theorem 4.5. *The total weight $E[W_k]$ of occupied urns after throwing k balls is given by*

$$E[W_k] = \sum_{i=1}^{m} W_i \cdot \left(1 - (1 - p_i)^k\right). \tag{4.5}$$

Proof. Consider the bivariate generating function

$$\Psi_\pi(x, y) = \sum_{j \geq 0} \sum_{k \geq 0} a_{j,k} \cdot x^j \cdot \frac{y^k}{k!}$$

where $a_{j,k}$ is the probability of reaching a set of urns having cumulated weight equal to j upon throwing k balls. Remark that such random allocations can be reinterpreted as sequences of m urns, each urn u_i containing either at non-empty set of balls (associated with a $x^{W_i}(e^{p_i y}-1)$ contribution) or no ball ($y^0 = 1$). Consequently the generating function $\Psi_\pi(x, y)$ can be reformulated as

$$\Psi_\pi(x, y) = \prod_{i=1}^{m} \left(1 + x^{W_i}\left(e^{p_i y} - 1\right)\right).$$

The generating function for the expectation of weight is then classically obtained through a partial derivative on x.

$$\begin{aligned} E[W_k] &= \left[\frac{y^k}{k!}\right] \frac{\partial \Psi_\pi(x, y)}{\partial z}(1, y) = \left[\frac{y^k}{k!}\right] e^{-y} \sum_{i=1}^{m} W_i \cdot \left(1 - e^{-y\, p_i}\right) \\ &= \sum_{i=1}^{m} \cdot \left(\left[\frac{y^k}{k!}\right] e^{y} - \left[\frac{y^k}{k!}\right] e^{y\,(1-p_i)} \right) = \sum_{i=1}^{m} W_i \cdot \left(1 - (1-p_i)^k\right) \end{aligned}$$

□

Remark that, upon setting $W_i = 1$, Equation 4.5 simplifies into $E[N_k]$ of urns reached by at least one ball (cf Hwang and Janson [17]), such that

$$E[N_k] = \sum_{i=1}^{m} \left(1 - (1-p_i)^k\right) = \sum_{i=1}^{m} \left(1 - e^{-p_i k}\right) + \mathcal{O}(1)$$

4.3.1. *Asymptotic estimates for the coverage.* Let us start from the formula

$$E[W_k] = \sum_{i=1}^{m} W_i \cdot \left(1 - (1-p_i)^k\right) = \sum_{i=1}^{m} W_i \cdot \left(1 - e^{k \cdot \log(1-p_i)}\right).$$

Since $p_i < 1$ for all $i \in [1, m]$, then one can use an approximation $\log(1 - p_i) = -p_i + \mathcal{O}(p_i^2)$ for large values of m, which can be be injected into E to obtain

$$E[W_k] = \sum_{i=1}^{m} W_i \cdot \left(1 - e^{k\left(-p_i + \mathcal{O}(p_i^2)\right)}\right).$$

If $k \cdot p_m \in o(1)$, then $k \cdot p_i \leq k \cdot p_m \in o(1)$ for all $i \in [1, m]$, and therefore $e^{k(-p_i + \mathcal{O}(p_i^2))} = 1 - kp_i + \mathcal{O}(kp_i^2)$, which gives

$$E[W_k] = \sum_{i=1}^{m} W_i \left(kp_i + \mathcal{O}(kp_i^2)\right) = k \sum_{i=1}^{m} W_i p_i + \mathcal{O}\left(k \sum_{i=1}^{m} W_i \cdot p_i^2\right). \tag{4.6}$$

In weighted languages that satisfy Condition **C1**, there exists $\beta > 1$ such that $p_i \in \mathcal{O}(\beta^{-n})$, for all $i \in [1, m]$. Consequently, for any $k \in o(\beta^n)$, one has

$$E[W_k] = k \sum_{i=1}^{m} W_i p_i \left(1 + \mathcal{O}(\beta^{-n})\right) = k \cdot \mu_{\pi,n} \cdot \alpha_{2,n} \left(1 + \mathcal{O}(\beta^{-n})\right).$$

5. Application to the statistical sampling of RNA

5.1. **Motivation.** Random generation has recently found a novel application in the *in silico* prediction of RNA folding. Namely a state-of-the-art method [8] for predicting the functional folding of a given RNA sequence uses a non-uniform random generation scheme [9]. This method aims at predicting the functional, or **native**, secondary structure of an RNA, a coarse-grain representation of the three-dimensional conformation. Based on the observation that the native structure is not necessarily that of lowest free-energy, Ding *et al* used a model initially proposed by Mc Caskill [18], and hypothesized a Boltzmann distribution based on the free-energy over the set of possible conformations. Their method generates a representative set of 1000 secondary structures

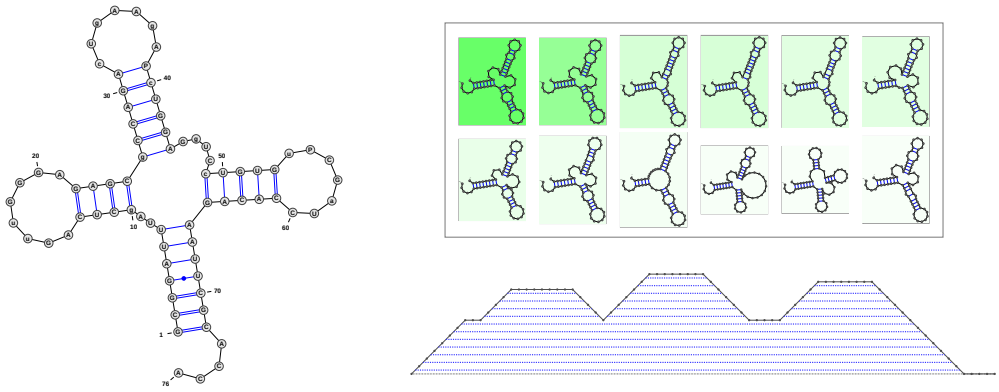

FIGURE 2. Secondary structure (Left) of a transfer RNA (tRNA) and its equivalent representation as a Motzkin walk (Bottom-right). Top-right: Typical picture of the Boltzmann ensemble, i.e. a set secondary structures compatible with the RNA sequence, colored according to their respective Boltzmann factor $e^{\frac{E_s}{RT}}$.

using a statistical sampling algorithm [9]. These structures are then clustered and a consensus structure is extracted. Considering this consensus led to a better sensibility/specificity tradeoff than previous approaches based on free-energy minimization [25].

However, given the variability in length and sequence composition of real RNAs, the 1000 structures criterion seems somewhat arbitrary and may lead to irreproducible observations in the context of highly variable observables. On the other hand, the sampled sets of structures might feature a large level of redundancy. Our theorems provide useful tools for a quantitative characterization of such situations.

5.2. **Statistical sampling of RNA secondary structures.** An RNA sequence can be encoded by a sequence of bases A, C, G and U where local compatibility rules (A↔U, A↔U, and G↔U) allow for a folding, i.e. a formation of chemical bounds between pairs of bases. The RNA secondary structure constitutes a restriction of all possible base-pairings, where each base is involved in at most one base-pairs with the additional constraint that the induced matching does not feature crossing interactions. A simplified energy model of Nussinov [19] assigns free-energies contributions E_b between -3.0 and -1.0 KCal.Mol^{-1} to each base-pairs b, depending on the number of hydrogen bonds involved in the interaction. The total free-energy $E_s = \sum_{b\in s} E_b$ of a secondary structure s is then inherited additively, and each secondary structure s is drawn with probability proportional to its **Boltzmann factor** $e^{\frac{E_s}{RT}}$ where R is the perfect gaz constant and T the temperature in Kelvin.

5.3. **Statistical sampling as a weighted generation.** Let us first remind that Motzkin words are well-parenthesized words featuring any number of dots characters $\bullet$. Let us define a **peak** as an occurrence of a motif $(\ )$, and a k-**plateau** as an occurrence of a motif $(\ \bullet^k\ )$, $k > 0$. Let $\theta \in \mathbb{N}$ be a parameter, then one defines secondary structures as *peakless* Motzkin words, or more generally as Motzkin words that are free of t-plateaux, for any $t < \theta$. The correspondence between coarse-grained conformations and Motzkin words is illustrated in Figure 2. Each pair of matching parentheses represents a base-pair, and the θ constant models steric constraints and is typically set to 1 in combinatorial studies [23] and to 3 in most RNA folding software. Through an adaptation of Viennot *et al* [22], secondary structures can be generated from a non-terminal S using rules

$$S \to (S_{\geq\theta})S \mid \bullet S \mid \varepsilon \qquad\qquad S_{\geq\theta} \to (S_{\geq\theta})S \mid \bullet S_{\geq\theta} \mid \bullet^\theta .$$

5.4. **Expected times for first collision and full collection.** Assuming a standard homopolymer model, in which any pair of base can bind, statistical sampling is equivalent to a weighted random

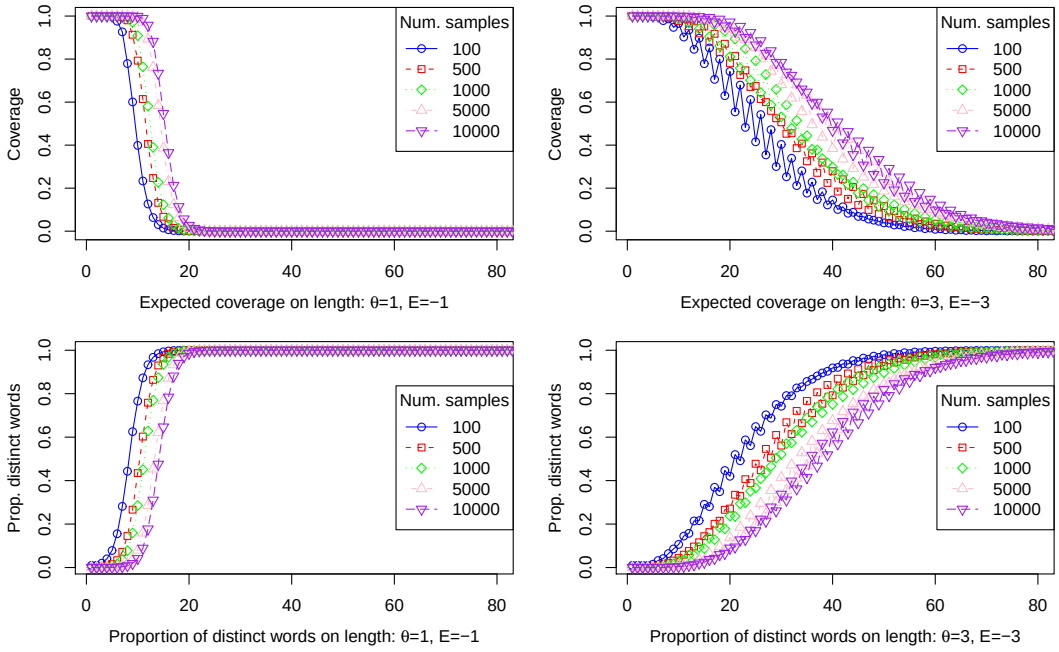

FIGURE 3. Expected coverage (Top) and proportion of distinct words (Bottom) within sampled set of words of various length, considering different values for θ and E the free-energy contribution of a base-pair.

generation, taking $w := e^{\frac{E}{RT}}$ as the weight of any base-pair b (e.g. any occurrence of a opening parenthesis). The resulting weighted generating function is then given by

$$\begin{aligned} S_{w,\theta}(z) =& \frac{1 - 2z + (w+1)z^2 - wz^{\theta+2} - \sqrt{\Delta_{w,\theta}}}{(1-z)2z^2} \\ \Delta_{w,\theta} :=& 1 - 4z + (6-2w)z^2 + 4(w-1)z^3 + (w-1)^2 z^4 \\ & - 2wz^{\theta+2} + 4wz^{\theta+3} - 2w(1+w)z^{\theta+4} + w^2 z^{2\theta+4}. \end{aligned}$$

Using our formulae, one can get estimates for the waiting times $E[B_{n,\theta,E}]$ and $E[C_{n,\theta,E}]$ for the first collision and full collection respectively, and observes the following behaviors

$$E[B_{n,1,-1}] \sim \frac{1.24 \cdot 1.54^n}{\sqrt[4]{n^3}} \qquad E[B_{n,3,-3}] \sim \frac{0.85 \cdot 1.105^n}{\sqrt[4]{n^3}}$$

$$\frac{0.64 \cdot 4.33^n}{n\sqrt{n}} \lesssim E[C_{n,1,-1}] \lesssim \frac{1.24 \cdot 4.33^n}{\sqrt{n}} \qquad \frac{0.065 \cdot 12.65^n}{n\sqrt{n}} \lesssim E[C_{n,3,-3}] \lesssim \frac{0.11 \cdot 12.65^n}{\sqrt{n}}$$

First one sees that the nature of these growths is unaffected by a change of weights and/or values of θ. This is not really surprising, since the grammar is strongly connected and therefore always gives rise to generating functions whose singularities are of square-root type [10]. However the exponential growth factor is strongly affected by these variations with practical consequences. For instance considering tRNAs ($n = 80$) and using our first order approximation gives a time of first collision of $\sim 4.7\,.\,10^{13}$ samples in the $(\theta = 1, E = -1)$ model, while only ~ 93.55 samples are required in the $(\theta = 3, E = -3)$ model for the first collision to occur.

5.5. **Collisions and coverage.** Finally let us address the coverage and number of distinct samples obtained by a random generation scenario. Remark that RNA secondary structures of length n with k plateaux are in bijection with Motzkin words of length $n - k\theta$ with k peaks/plateaux, where the

bijection simply consists in removing the first θ horizontal steps of each plateau in every secondary structure. Let us further remind that Dyck words with k peaks and $2i$ letters are counted by the Naranaya numbers $\mathcal{N}(i,k)$, and that Motzkin words are obtained by inserting some dots within a Dyck word. It follows that the number $s_{n,k,i,\theta}$ of secondary structures of length n featuring i plateaux and $k \geq i$ base-pairs is such that

$$s_{n,k,i,\theta} = \mathcal{N}(i,k)\binom{n-\theta k}{n-2i-\theta k} = \frac{1}{i}\binom{i}{k}\binom{i}{k-1}\binom{n-\theta k}{n-2i-\theta k} \tag{5.1}$$

Using the above formula, one can compute exactly in polynomial time the expected coverage from Theorem 3.4 and the proportion of distinct samples from Theorem 3.3, and one obtains the results summarized in Figure 3. Interestingly Figure 3 shows that the inevitable decay of the coverage can be delayed by free-energies contributions of large absolute values. For instance a sampled set of 1000 structures still achieves a 50% coverage for RNAs of length < 30 for a free-energy contribution in the $(\theta = 3, E = -3)$ model while yielding a negligible coverage in the $(\theta = 1, E = -1)$ model. This suggests that, for highly stable RNAs (having low free-energy) of modest size, the 1000 structure criterion might be sufficient. Also a symmetry of the coverage and proportion can be observed, although the amplitude of the oscillations for $\theta = 3$ seem to have less of an impact on the proportion of distinct words than on their coverage.

6. Conclusion and perspectives

In this article, we investigated the redundancy of random sets of words of context-free languages drawn with respect to a weighted distribution. Using a random allocation model we derived exact and/or asymptotic equivalent forms for: the expected numbers of generations prior to the first collision and full collection, the average proportion of distinct words within a sampled set of k words and its cumulated probability. Interestingly, the second moment of the probability distribution both appears in the asymptotic behaviors of the first collision and the expected coverage. We applied these theorems to analyze the output of a statistical sampling algorithm used to predict the functional folding of RNA molecules. We showed that, although the time of first collision is exponential on the length of the RNA, its exponential factor strongly depends on the free-energy contribution of base-pairs, and may still allow for frequent collisions for RNAs of small – yet relevant – lengths.

Future directions for this work first include a better characterization of the full collection waiting time. Namely we showed that, unsurprisingly, the waiting time is dominated by the overall (exponential) weight but obtained lower and upper that are still separated by a $\Theta(n)$ factor. A possible direction for a tighter bound resides in algebraic manipulations of Harmonic numbers coupled with additional assumptions on the distribution of weights (i.e. distribution of symbols), for which local limit theorems are known to hold under certain hypotheses. Also we may refine our analysis of RNA statistical sampling, using more sophisticated – yet still context-free – grammars in order to accommodate more realistic models for the free-energy.

Acknowledgements

The authors wish to thank the organizers of the GASCOM'08 conference in Bibbiena, Italy where the present collaboration started. The present work was funded the *Agence Nationale de la Recherche* through the BOOLE NT09_432755 (DG) and the GAMMA 07-2_195422 (YP) programs.

References

1. F. Bassino, J. David, and C. Nicaud, *On the average complexity of Moore's state minimization algorithm*, 26th International Symposium on Theoretical Aspects of Computer Science (STACS 2009) (Dagstuhl, Germany), Leibniz International Proceedings in Informatics (LIPIcs), vol. 3, 2009, pp. 123–134.
2. P. Berenbrink and T. Sauerwald, *The weighted coupon collector's problem and applications*, 15th International Computing and Combinatorics Conference (COCOON'10), 2009.

3. O. Bodini and Y. Ponty, *Multi-dimensional Boltzmann sampling of languages*, Proceedings of AOFA'10 (Vienna), June 2010.
4. S. Brlek, E. Pergola, and O. Roques, *Non uniform random generation of generalized Motzkin paths*, Acta Informatica **42** (2006), no. 8, 603–616.
5. B. Canou and A. Darrasse, *Fast and sound random generation for automated testing and benchmarking in objective Caml*, ML '09: Proceedings of the 2009 ACM SIGPLAN workshop on ML (New York, NY, USA), 2009, pp. 61–70.
6. A. Denise, M.-C. Gaudel, S.-D. Gouraud, R. Lassaigne, and S. Peyronnet, *Uniform random sampling of traces in very large models*, First ACM International Workshop on Random Testing (ISSTA), 2006, pp. 10–19.
7. A. Denise, O. Roques, and M. Termier, *Random generation of words of context-free languagesaccording to the frequencies of letters*, Mathematics and Computer Science: Algorithms,Trees, Combinatorics and probabilities (D. Gardy and A. Mokkadem, eds.), Trends in Mathematics, Birkhaüser, 2000, pp. 113–125.
8. Y. Ding, C. Y. Chan, and C. E. Lawrence, *RNA secondary structure prediction by centroids in a boltzmann weighted ensemble*, RNA **11** (2005), 1157–1166.
9. Y. Ding and E. Lawrence, *A statistical sampling algorithm for RNA secondary structure prediction*, Nucleic Acids Research **31** (2003), no. 24, 7280–7301.
10. M. Drmota, *Systems of functional equations*, Random Struct. Alg. **10** (1997), 103–124.
11. P. Duchon, P. Flajolet, G. Louchard, and G. Schaeffer, *Boltzmann samplers for the random generation of combinatorial structures*, Combinatorics, Probablity, and Computing **13** (2004), no. 4–5, 577–625, Special issue on Analysis of Algorithms.
12. P. Flajolet, *Analytic models and ambiguity of context-free languages*, Theor. Comput. Sci. **49** (1987), no. 2-3, 283–309.
13. P. Flajolet, D. Gardy, and L. Thimonier, *Birthday paradox, coupon collectors, caching algorithms and self-organizing search*, Discrete Appl. Math. **39** (1992), no. 3, 207–229.
14. P. Flajolet and R. Sedgewick, *Analytic combinatorics*, Cambridge University Press, 2009.
15. P. Flajolet, P. Zimmermann, and B. Van Cutsem, *Calculus for the random generation of labelled combinatorial structures*, Theoretical Computer Science **132** (1994), 1–35.
16. D. Gardy, *Occupancy urn models in the analysis of algorithms*, Journal of Statistical Planning and Inference **101** (2002), no. 1-2, 95 – 105.
17. H.-K. Hwang and S. Janson, *Local limit theorems for finite and infinite urn models.*, Ann. Probab. **36** (2008), no. 3, 992–1022.
18. J.S. McCaskill, *The equilibrium partition function and base pair binding probabilities for RNA secondary structure*, Biopolymers **29** (1990), 1105–1119.
19. R. Nussinov and A.B. Jacobson, *Fast algorithm for predicting the secondary structure of single-stranded rna*, Proc Natl Acad Sci U S A **77** (1980), 6903–13.
20. Y. Ponty, *Non-redundant random generation from weighted context-free languages*, Proceedings of GASCOM'08, June 2008, p. 10 pp.
21. Y. Ponty, M. Termier, and A. Denise, *GenRGenS: Software for generating random genomic sequences and structures*, Bioinformatics **22** (2006), no. 12, 1534–1535.
22. M. Vauchaussade de Chaumont and G. Viennot, *Polynômes orthogonaux et problèmes d'énumération en biologie moléculaire*, Séminaire Lotharingien de Combinatoire (1983).
23. M. S. Waterman, *Secondary structure of single stranded nucleic acids*, Advances in Mathematics Supplementary Studies **1** (1978), no. 1, 167–212.
24. H. S. Wilf, *A unified setting for sequencing, ranking, and selection algorithms for combinatorial objects*, Advances in Mathematics **24** (1977), 281–291.
25. M. Zuker and P. Stiegler, *Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information*, Nucleic Acids Res **9** (1981), 133–148.

Appendix

Proof. (Theorem 4.1) From [13], the waiting time for the first birthday can be expressed as

$$E(B) = \int_0^{+\infty} \lambda(t)e^{-t}dt, \qquad with \qquad \lambda(t) = \prod_{i=1}^{m}(1+p_it).$$

Let us approximate this integral under the conditions of the theorem. We cut the integral at τ, and independently consider the part from 0 to τ, which we expect will be dominating, and the part from τ to $+\infty$, which should give rise to a negligible contribution.

Let us first approximating the integral $\int_0^{\tau} \lambda(t)e^{-t}dt$. Consider

$$\psi(t) = \log\lambda(t) - t = \sum_{i=1}^{m} \log(1+p_it) - t.$$

Then $E(B) = \int_0^{+\infty} e^{\psi(t)}dt$. Let us consider a positive real value $t < \tau$ such that any value p_it is uniformly bounded by some $A < 1$, then $\log(1+p_it) = p_it - p_i^2t/2 + \mathcal{O}(p_i^3t^3)$, where the bound implied in the $\mathcal{O}(\cdot)$ term is uniform in p_it. Summing over the whole distribution gives

$$\psi(t) = -\left(\sum_i p_i^2\right) t^2/2 + \mathcal{O}\left(\sum_i p_i^3t^3\right)$$

then $\psi(t) = -\alpha_2t^2/2 + \mathcal{O}(\alpha_3t^3) = -\alpha_2t^2/2 + \mathcal{O}(\alpha^3\tau^3)$. Plugging this into the integral gives

$$\int_0^{\tau} e^{-\alpha_2t^2/2+\mathcal{O}(\alpha^3\tau^3)}dt = \int_0^{\tau} e^{-\alpha_2t^2/2}dt.(1+\mathcal{O}(\alpha_3\tau^3)).$$

This last integral is computed by a change of variable $u = t\sqrt{\alpha_2}$. Approximating with a Gaussian integral $\int_0^{+\infty} e^{-u^2/2}du = \sqrt{\pi/2}$ finally gives

$$\int_0^{\tau} \lambda(t)e^{-t}dt = \sqrt{\frac{\pi}{2\alpha_2}}\left(1+\mathcal{O}\left(e^{-\tau^2\alpha_2/2}\right)+\mathcal{O}\left(\alpha_3\tau^3\right)\right).$$

Of course, the validity of this expansion requires that

- The error terms in the above equation are $o(1)$: This follows from our assumptions on τ, reminding that $\alpha_3\tau^3 \to 0$ (Condition (C)) and $\alpha_2\tau^2 \to \infty$ (Condition (B)).
- Each of the terms p_it is uniformly bounded by some $A < 1$: Since p_m is the greatest probability, then it suffices that $p_m\tau \leq A < 1$ (Condition (A)).

Let us bound the value of the remainder $\int_{\tau}^{+\infty} \lambda(t)e^{-t}dt$. We factor out the term $\lambda(\tau)e^{-\tau} = e^{\psi(\tau)}$, which we expect to be dominant. The remaining term is

$$\begin{aligned}
\int_{\tau}^{+\infty} \frac{\lambda(t)}{\lambda(\tau)}e^{\tau-t}dt &= \int_0^{+\infty} \frac{\lambda(\tau+s)}{\lambda(\tau)}e^{-s}ds \\
&= \int_0^{+\infty} \prod_{i=1}^{m}\left(1+\frac{p_i}{1+p_i\tau}\,s\right)e^{-s}ds \\
&= \int_{\tau}^{+\infty} e^{\sum_{i=1}^{m}\log\left(1+\frac{p_i}{1+p_i\tau}\,s\right)-s}ds
\end{aligned}$$

Now, for any positive x, $\log(1+x) \leq x$ which gives a bound

$$\begin{aligned}
\sum_{i=1}^{m}\log\left(1+\frac{p_i}{1+p_i\tau}\,s\right) - s &\leq \left(\sum_{i=1}^{m}\frac{p_i}{1+p_i\tau}\,s\right) - s \\
&\leq \sum_{i=1}^{m}\left(\frac{p_i}{1+p_i\tau} - p_i\right)s = -sB(\tau),
\end{aligned}$$

with $B(\tau) = \sum_{i=1}^{m} \left(\frac{-p_i}{1+p_i\tau} + p_i \right) = \sum_i \frac{p_i^2 \tau}{1+p_i\tau}$. It follows that

$$\int_0^{+\infty} \frac{\lambda(s+\tau)}{\lambda(\tau)} e^{-s} ds \leq \int_0^{+\infty} e^{-B(\tau)s} ds = \frac{1}{B(\tau)}$$

and finally

$$\int_\tau^{+\infty} \lambda(t) e^{-t} dt \leq \frac{\lambda(\tau)\, e^{-\tau}}{B(\tau)}.$$

Let us consider the order of $B(\tau)$. We easily check that, for each i,

$$0 < 1 - p_i\tau < \frac{1}{1+p_i\tau} < 1,$$

thus

$$0 < p_i^2\tau - p_i^3\tau^2 < \frac{p_i^2\tau}{1+p_i\tau} < p_i^2\tau,$$

which gives bounds on $B(\tau)$ as

$$\alpha_2\tau - \alpha_3\tau^2 < B(\tau) < \alpha_2\tau.$$

Finally, we can bound the error term. In order to conclude, we need to show that

$$\lambda(\tau)e^{-\tau}/B(\tau) = o(1/\sqrt{\alpha_2}).$$

First rewrite the last condition as $e^{-\alpha_2\tau^2/2+\mathcal{O}(\alpha_3\tau^3)} = o(B(\tau)/\sqrt{\alpha_2})$, taking advantage of $\lambda(\tau) = e^{\tau-\alpha^2\tau^2/2+O(\alpha_3\tau^3)}$. Assume that we have chosen τ such that $\alpha_2\tau \to +\infty$ and $\alpha_3\tau^3 \to 0$; then $B(\tau)$ has exact order $\alpha_2\tau$ and the condition collapses to $e^{-\alpha_2\tau^2/2+O(\alpha_3\tau^3)} = o(\tau\sqrt{\alpha_2})$, which is trivial. □

Proof. (Theorem 4.2) Let us first remind that the **exponential order** [14] of a sequence f_n, is a simple exponential function K^n such that

$$\lim_{n\to\infty} \sup |f_n|^{1/n} = K.$$

Following notations of the Flajolet/Sedgewick's book [14], we make use of the *bowtie* notation, and write $f_n \bowtie K^n$ if f_n has exponential order K^n. It is a classic result [14, Theorem IV.7] that the dominant singularity ρ of a generating function determines the exponential order of its coefficients c_n, namely through $c_n \bowtie \rho^{-n}$.

Since ${\rho_\pi}^k < \rho_{\pi^k}$ holds for any $k > 1$ and $\pi_0 > \mathbf{1}$ (Condition **C3**), then it follows that

$$s_{n,k} := \sqrt[k]{\mu_{\pi_0^k,n}} \bowtie \left(\sqrt[k]{\rho_{\pi_0^k}} \right)^{-n} \quad \text{and} \quad \sqrt[k]{\rho_{\pi_0^k}} > \rho_{\pi_0} \tag{6.1}$$

for π_0 any vector of weights strictly larger than 1, and ρ_{π_0} the dominant singularity of $L_\pi(z)$.

This result generalizes to any pair $(a, b) \in \mathbb{R}^2$ of numbers such that $1 < a < b$. Indeed, upon taking $\pi_0 = \pi^a$ and $k = b/a$ in the above equation, one has $s_{n,a} \bowtie \left(\sqrt[a]{\rho_{\pi^a}} \right)^{-n}$, $s_{n,b} \bowtie \left(\sqrt[b]{\rho_{\pi^b}} \right)^{-n}$, and it follows from Condition **C3** that

$$\sqrt[a]{\rho_{\pi^a}} < \sqrt[b]{\rho_{\pi^b}}. \tag{6.2}$$

Consequently, for any $1 < a < b$, $s_{n,a}$ grows exponentially faster than $s_{n,b}$, and one can use such a hierarchy to *squeeze* τ_n^{-1} between $\sqrt{\alpha_{2,n}}$ and $\sqrt[3]{\alpha_{3,n}}$.

Namely let us consider

$$\tau_n := \frac{1}{\sqrt[k]{\alpha_{k,n}}}$$

for some $k \in \mathbb{Q}$ such that $2 < k < 3$. Then we have

$$\sqrt{\alpha_{2,n}} \cdot \tau_n = \sqrt{\frac{\mu_{\pi^2,n}}{{\mu_{\pi,n}}^2}} \sqrt[k]{\frac{{\mu_{\pi,n}}^k}{\mu_{\pi^k,n}}} = \frac{s_{n,2}}{s_{n,k}} \bowtie \left(\frac{\sqrt[k]{\rho_{\pi^k}}}{\sqrt{\rho_{\pi^2}}} \right)^n$$

and it follows from $\sqrt{\rho_{\pi^2}} < \sqrt[k]{\rho_{\pi^k}}$ that

$$\lim_{n\to\infty} \sqrt{\alpha_{2,n}} \cdot \tau_n = +\infty$$

and consequently Condition (B) is satisfied by our candidate τ_n.

Reciprocally for Condition (C), one has

$$\sqrt[3]{\alpha_{3,n}} \cdot \tau_n = \sqrt[3]{\frac{\mu_{\pi^3,n}}{{\mu_{\pi,n}}^3}} \sqrt[k]{\frac{{\mu_{\pi,n}}^k}{\mu_{\pi^k,n}}} = \frac{s_{n,3}}{s_{n,k}} \bowtie \left(\frac{\sqrt[k]{\rho_{\pi^k}}}{\sqrt[3]{\rho_{\pi^3}}} \right)^n$$

and, since $\sqrt[3]{\rho_{\pi^3}} > \sqrt[k]{\rho_{\pi^k}}$, then

$$\lim_{n\to\infty} \sqrt[3]{\alpha_{3,n}} \cdot \tau_n = 0.$$

Condition (A) is also satisfied by τ_n upon observing that

$$p^{\triangle}_{\pi,n} \cdot \tau_n = \frac{W^{\triangle}_{\pi,n}}{\mu_{\pi,n}} \sqrt[k]{\frac{{\mu_{\pi,n}}^k}{\mu_{\pi^k,n}}} = \frac{W^{\triangle}_{\pi,n}}{s_{n,k}}$$

where $W^{\triangle}_{\pi,n}$ is the weight of the heaviest (i.e. most probable) word $w^{\triangle} \in \mathcal{L}_n$. This word is also contributing to $\mu_{\pi^k,n} = \sum_{w\in\mathcal{L}_n} \pi(w)^k$ and therefore

$$s_{n,k} = \sqrt[k]{\mu_{\pi^k,n}} = \sqrt[k]{{W^{\triangle}_{\pi,n}}^k + \sum_{\substack{w\in\mathcal{L}_n \\ w\neq w^{\triangle}}} \pi(w)^k} > W^{\triangle}_{\pi,n}$$

which suffices to prove that Condition (A) is satisfied. Consequently, the preconditions of Theorem 4.1 are satisfied by any weighted distribution. □

Laboratoire PRiSM.CNRS UMR 8144 and Université de Versailles St-Quentin en Yvelines, 45 Av. des États-Unis, 78035 Versailles, France

E-mail address: Daniele.Gardy@prism.uvsq.fr

Laboratoire d'Informatique de l'École Polytechnique (LIX), CNRS UMR 7161/AMIB INRIA. École Polytechnique, 91128 Palaiseau, France

E-mail address: yann.ponty@lix.polytechnique.fr